deFUME: Dynamic exploration of functional metagenomic sequencing data

van der Helm, Eric; Geertz-Hansen, Henrik Marcus; Genee, Hans Jasper; Malla, Sailesh; Sommer, Morten Otto Alexander

Published in:
BMC Research Notes

Link to article, DOI:
10.1186/s13104-015-1281-y

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

van der Helm *et al. BMC Res Notes (2015) 8:328*
DOI 10.1186/s13104-015-1281-y

BMC Research Notes

TECHNICAL NOTE

Open Access

deFUME: Dynamic exploration of functional metagenomic sequencing data

Eric van der Helm[1*], Henrik Marcus Geertz-Hansen[1,2,3], Hans Jasper Genee[1], Sailesh Malla[1] and Morten Otto Alexander Sommer[1,4]

Abstract

Background: Functional metagenomic selections represent a powerful technique that is widely applied for identification of novel genes from complex metagenomic sources. However, whereas hundreds to thousands of clones can be easily generated and sequenced over a few days of experiments, analyzing the data is time consuming and constitutes a major bottleneck for experimental researchers in the field.

Findings: Here we present the deFUME web server, an easy-to-use web-based interface for processing, annotation and visualization of functional metagenomics sequencing data, tailored to meet the requirements of non-bioinformaticians. The web-server integrates multiple analysis steps into one single workflow: read assembly, open reading frame prediction, and annotation with BLAST, InterPro and GO classifiers. Analysis results are visualized in an online dynamic web-interface.

Conclusion: The deFUME webserver provides a fast track from raw sequence to a comprehensive visual data overview that facilitates effortless inspection of gene function, clustering and distribution. The webserver is available at cbs.dtu.dk/services/deFUME/and the source code is distributed at github.com/EvdH0/deFUME.

Keywords: Functional metagenomics, Web services, Visualization, Sequence analysis, Gene ontology

Findings

Background

Functional selection represents a powerful technique for discovery and functionally validated annotation of genes [1, 2]. The technique relies on the expression of randomly cloned genomic or metagenomic DNA, typically as short (1–3 kb) fragments of DNA into an expression vector. The expression library is subsequently transformed into a suitable host where the desired functionality can be selected. DNA inserts from clones exhibiting the desired phenotype can be sequenced, enabling functional isolation of novel genes. The approach has been applied for identification of genes from complex metagenomic sources and examples include DNA polymerases [3], antibiotics resistance genes [1, 2, 4, 5], xenobiotic degradative enzymes [6], and more [7].

Most commonly, inserts of 1–3 kb are sequenced either by next generation sequencing [4, 8] or by conventional bi-directional Sanger sequencing, followed by base-calling, quality trimming of reads and assembly of reads into contigs. Finally data analysis is performed including BLAST searches and other functional annotations. Whereas hundreds to thousands of clones can be easily generated over a few days of experiments, analyzing the data is time consuming and constitutes a major bottleneck for experimental researchers in the field. To address this challenge we developed deFUME; an easy-to-use web server that automatically processes and annotates large amounts of sequencing data obtained through functional selections.

Implementation

As input, deFUME accepts nucleotide sequences generated from next generation sequencing technologies as well as raw reads from Sanger sequencing. Sequences can be uploaded via the submission page either as

*Correspondence: evand@biosustain.dtu.dk
[1] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2970 Hørsholm, Denmark
Full list of author information is available at the end of the article

preassembled projects, when using next generation sequencing data in. FASTA format, or as raw Sanger sequencing chromatograms (.ab1 format). For Sanger sequencing chromatograms, base-calling and assembly of the resulting reads is carried out with Phred [9] (default parameters) and Phrap [10] (using the non-default parameters -minscore 25, -trim_score 20, and -min_match 20 to reflect medium stringency cutoffs in order to ensure good quality assemblies of Sanger reads) respectively. Prior to assembly the vector sequence is optionally masked in the reads using Cross_match (using -minmatch 12 -minscore 20) [10] (Additional file 1: Figure S1).

Open reading frames (ORFs) are predicted using MetaGeneMark [11] (default parameters) from the resulting assembly (generated by Phrap or user input). The translated ORFs are aligned to the nr protein database using BLASTp (reporting only the 25 most significant hits with a minimal E value of 0.001) [12] and submitted to InterPro [13] (default parameters). The InterPro database contains signatures of known proteins families that can be queried to functionally characterize new sequences. To ensure using the most recent databases, InterPro is accessed using the simple object access protocol (SOAP) via InterProScan 5 [14].

The deFUME output page (Fig. 1a) is an interactive table showing the sequencing reads, the assembled contigs, predicted ORFs, BLASTp hits and InterPro functional data. The user can highlight hits and filter (Fig. 1b) the data by parameters such as BLAST E value, specific GO terms and removal of hypothetical protein homologs. Finally, the user can export selected contigs in FASTA, Genbank and CSV file formats.

The back-end pipeline is written in the programming language Perl (version 5.8.7) and PHP5 and the front-end visualization in a combination of JavaScript and HTML5 using the D3js, jqGrid and jQuery packages. The access to the webserver is free and unlimited for all academic users with a maximal data upload time of 2 min per job and a maximal job runtime of 24 h. The source code is freely available at https://github.com/EvdH0/deFUME for continuous improvement and development by the community.

Results and discussion

To demonstrate and test the deFUME web server, we analyzed Sanger sequencing data derived from a functional metagenomics selection for genes conferring tolerance of *E. coli* to high levels of lysine. Briefly, metagenomic DNA derived from cow fecal matter was mechanically sheared and subsequently cloned and expressed in *E. coli.* The resulting cell library was subjected to high lysine levels on Luria–Bertani agar plates (Additional file 1). The inserts of 80 individual colonies tolerant to high lysine levels were sequenced. The resulting 160 raw Sanger sequencing chromatograms were submitted as.ab1 files to the deFUME web server. In less than 2 h, all reads were trimmed and assembled, resulting in 69 unique contigs, 117 predicted ORFs, 134 GO annotations and 622 InterPro functional predictions. A screenshot of the output page is shown in Fig. 1, displaying one of the 69 inserts. In this particular insert, deFUME predicted three ORFs with a sequence coverage (calculated by dividing the protein sequence length of the predicted ORF by the length of the best BLASTp hit) of 12, 100 and 90%. Together with an E value of 4^{-52} this indicates that ORF 2 likely

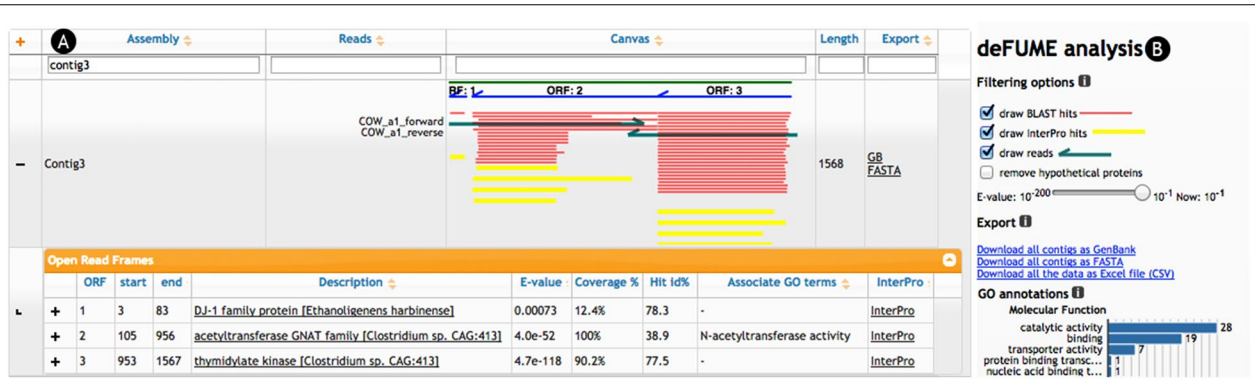

Fig. 1 The deFUME output page. **a** Screenshot from the deFUME interactive output showing two Sanger reads (*dark green*) assembled into a 1,568 bp contig (*green*) by Phrap. deFUME annotated three open reading frames; ORF1 is annotated by BLASTp as a DJ-1 family protein, ORF2 as an acetyltransferase and ORF3 as a thymidylate kinase. The "Coverage %" *column* shows that only ORF2 is a complete gene, indicating the phenotypic activity from this clone likely arises from ORF2. All ORFs contain multiple Interpro hits (*yellow*) that can be inspected in detail by clicking on the "Interpro" link which launches the native Interpro webpage. **b** deFUME analysis toolbox. The data can be filtered and manipulated by turning on and off different levels of metadata, filter on E value and filter on specific GO terms.

encodes the mechanism for lysine tolerance. The translated ORF is annotated by BLASTp as an "acetyltransferase GNAT family" protein catalyzing the acetylation of the nitrogen group of lysine, thus providing tolerance to high levels of lysine. The example demonstrates the ability of deFUME to accelerate the overall process of going from experimental raw data to functional annotation.

Conclusions

deFUME is the first web server to integrate all steps from sequencing read assembly to a comprehensive visual output for functional annotation of metagenomic insert libraries. Additionally, deFUME reduces the hands-on time required for analysis compared to packages like CLC Main Workbench (CLC Bio, Aarhus, Denmark) and Mobyle [15], where the user has to transfer the intermediate data from one tool to the other. Furthermore, the data integration and visualization of deFUME is substantially more advanced compared to the current state-of-art by providing interactive exploration of heterogeneous data.

Availability and requirements

Project name: deFUME.

Project home page: http://www.cbs.dtu.dk/services/deFUME.

Operating system(s): Platform independent.

Programming language: Perl, PHP, Javascript, HTML5.

Other requirements: Browser supporting HTML5.

License: Creative Commons BY 2.0

Any restrictions to use by non-academics: no.

Additional file

Additional file 1. Supplementary information.

Authors' contributions

EvdH, HJG, HMGH and MOAS conceived the study. EvdH and HMGH wrote the software. SM performed the experimental work. All authors read and approved the final manuscript.

Author details

[1] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2970 Hørsholm, Denmark. [2] Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, 2800 Lyngby, Denmark. [3] Novozymes A/S, 2880 Bagsværd, Denmark. [4] Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark.

Acknowledgements

The authors thank Dionísio S. Paiva for providing the cow fecal library and Hans H. Stærfeldt for advice on the web server implementation.

Compliance with ethical guidelines

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the European Union Seventh Framework Programme—ITN [FP7/2012/317058 to EvdH]; Novozymes A/S [HJG, HMGH]; the European Union Seventh Framework Programme (FP7-KBBE-2013-7-single-stage) under grant agreement no. 613745, Promys [MOAS, HJG, EvdH] and the Novo Nordisk Foundation.

Received: 19 May 2015 Accepted: 15 July 2015
Published online: 31 July 2015

References

1. Riesenfeld CS, Goodman RM, Handelsman J (2004) Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. Environ Microbiol 6:981–989
2. Sommer MOA, Dantas G, Church GM (2009) Functional characterization of the antibiotic resistance reservoir in the human microflora. Science 325:1128–1131
3. Simon C, Herath J, Rockstroh S, Daniel R (2009) Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. Appl Environ Microbiol 75:2964–2968
4. Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MOA, Dantas G (2012) The shared antibiotic resistome of soil bacteria and human pathogens. Science 337(6098):1107–1111
5. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N et al (2014) Bacterial phylogeny structures soil resistomes across habitats. Nature 509:612–616
6. Ono A, Miyazaki R, Sota M, Ohtsubo Y, Nagata Y, Tsuda M (2007) Isolation and characterization of naphthalene-catabolic genes and plasmids from oil-contaminated soil by using two cultivation-independent approaches. Appl Microbiol Biotechnol 74:501–510
7. Uchiyama T, Miyazaki K (2009) Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr Opin Biotechnol 20:616–622
8. Udikovic-Kolic N, Wichmann F, Broderick NA, Handelsman J (2014) Bloom of resident antibiotic-resistant bacteria in soil following manure fertilization. Proc Natl Acad Sci 111:15202–15207
9. Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res 8:175–185
10. Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202
11. Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38:e132
12. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K et al (2009) BLAST+ : architecture and applications. BMC Bioinform 10:421
13. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A et al (2012) InterPro in 2011: New developments in the family and domain prediction database. Nucleic Acids Res 40(November 2011):306–312
14. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C et al (2014) InterProScan 5: Genome-scale protein function classification. Bioinformatics 30:1236–1240
15. Néron B, Ménager H, Maufrais C, Joly N, Maupetit J, Letort S et al (2009) Mobyle: a new full web bioinformatics framework. Bioinformatics 25:3005–3011